



دانشکده مهندسی  
کامپیوتر و فناوری اطلاعات

دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

دانشکده مهندسی کامپیوتر



دانشگاه صنعتی امیرکبیر  
( پلی تکنیک تهران )

# داده کاوی

(بهار ۱۴۰۱)

## تمرین اول

محمد چوپان ۹۸۳۱۱۲۵

## بخش تئوری :

### سوال اول :

#### سوال اول

به سوالات زیر پاسخ دهید.

الف) داده‌ی پرت<sup>۱</sup> با نویز<sup>۲</sup> را با یکدیگر مقایسه کنید.

ب) یک سناریو بیان کنید که در آن داده‌های پرت برای ما مفید هستند و اطلاعات ارزشمندی از آن دریافت می‌کنیم.

ج) مشخص کنید که آیا یک نویز می‌تواند داده‌ی پرت باشد یا خیر؟

### پاسخ:

#### الف :

داده پرت و داده نویز دو مفهوم متفاوت هستند. داده پرت به داده‌ای اطلاق می‌شود که اشتباهاتی در آن وجود دارد ولی به صورت تصادفی نیست و از خطای اندازه‌گیری و یا مشکلات در جمع‌آوری داده‌ها ناشی می‌شود. از طرف دیگر، داده نویز به داده‌ای اطلاق می‌شود که شامل اشتباهات تصادفی است که به دلیل نویز در داده‌ها به وجود می‌آید.

برای مقایسه داده پرت و داده نویز، می‌توانیم از یک معیار به نام MSE یا میانگین مربعات خطا استفاده کنیم. این معیار میزان اختلاف بین داده‌های پیش‌بینی شده و داده‌های واقعی را در قالب یک مقدار عددی برمی‌گرداند. به عبارت دیگر، MSE نشان می‌دهد که در صورت استفاده از یک الگوریتم پیش‌بینی، چقدر خطای میانگین میان داده پیش‌بینی شده و داده واقعی است.

اگر داده پرت و داده نویز را با یکدیگر مقایسه کنیم، MSE برای داده پرت کمتر از MSE داده نویز خواهد بود، چرا که داده پرت حداقل از یک الگوریتم مشخص استفاده کرده است و اختلاف‌های آن نسبتاً ثابت و پیش‌بینی‌پذیر هستند. از طرف دیگر، داده نویز شامل خطاهای تصادفی است که به سختی قابل پیش‌بینی و کنترل هستند.

#### ب :

یکی از مثال‌هایی که در آن داده‌های پرت ارزشمند هستند، ممکن است این باشد که در حوزه طب، داده‌های بیماری مورد بررسی قرار می‌گیرد. در بعضی موارد، داده‌های بیماری ممکن است شامل خطاهای پرت باشد که از خطای اندازه‌گیری و یا عوامل دیگر ناشی می‌شود، اما این داده‌ها همچنان ارزشمند هستند، زیرا ممکن است حاوی اطلاعات مفید و مخفی باشد.

برای مثال، فرض کنید یک محقق در حوزه طب قصد دارد تاثیر یک درمان جدید برای یک بیماری خاص را بررسی کند. با جمع‌آوری داده‌های بیماری از بیماران مختلف، ممکن است برخی داده‌ها شامل خطاهای پرت باشند. اما اگر با استفاده از الگوریتم‌های داده کاوی، این خطاها شناسایی و از داده‌های مورد بررسی حذف شوند، ممکن است محقق اطلاعات جالبی از نحوه عملکرد درمان جدید بر روی بیماران باشند. در این حالت، داده‌های پرت از ما مفید هستند ولی با کاوش در این داده‌ها می‌توان از اطلاعات ارزشمندی به دست آورد. بنابراین، در این حالت، داده‌های پرت به دلیل اینکه حاوی اطلاعات مفیدی هستند، ارزشمند می‌شوند و حذف آنها می‌تواند باعث از دست رفتن اطلاعاتی ارزشمند شود.

بله، یک نویز می‌تواند داده‌ای پرت باشد. در واقع، نویز به عنوان یک عامل مخرب می‌تواند باعث ایجاد خطا در داده شود و باعث شود که داده‌ها پرت شوند. به طور کلی، نویز به دلیل تغییر دادن مقدار داده‌های اصلی و ایجاد اختلال در آنها، ممکن است باعث شود که داده‌ها پرت شوند و باعث کاهش دقت و قابلیت اطمینان در مدل‌سازی و پیش‌بینی شود. به عنوان مثال، در داده‌های سنجش دما، نویز می‌تواند ناشی از تقلب در داده‌های اندازه‌گیری، مشکل در دستگاه اندازه‌گیری یا نویزهای محیطی باشد. همچنین، در داده‌های صوتی، نویز می‌تواند ناشی از نویزهای محیطی، مشکلات در دستگاه ضبط صوت یا نویزهای ناشی از فرآیند فشرده‌سازی صوت باشد. بنابراین، نویز می‌تواند یکی از عواملی باشد که باعث ایجاد داده‌های پرت شود و می‌تواند اطلاعات ارزشمند در داده‌ها را به شدت کاهش دهد.

## سوال دوم :

در حوزه‌ی داده‌کاوی، انبار داده<sup>۳</sup> چیست و چه تفاوت و شباهتی با پایگاه داده<sup>۴</sup> دارد؟

## پاسخ:

انبار داده (Data Warehouse) به مجموعه‌ای از داده‌ها اطلاق می‌شود که برای پشتیبانی از تصمیم‌گیری‌های کسب و کار استفاده می‌شود. یک انبار داده معمولاً شامل داده‌هایی است که از منابع مختلفی جمع‌آوری شده‌اند، مانند پایگاه داده‌های تعاملی کاربری، سیستم‌های مدیریت منابع سازمانی (ERP)، سیستم‌های پشتیبانی از تصمیم‌گیری (DSS) و سایر منابع داده. هدف انبار داده‌ها برای کسب و کارها، تجمیع و ادغام داده‌های مختلف در یک محیط مرکزی است که به کاربران امکان می‌دهد به سرعت و با دقت بالا از آن استفاده کنند. به این ترتیب، انبار داده‌ها قابلیت ارائه داده‌های مرتبط، موجودیت‌های اطلاعاتی، نمودارها، گزارشات، و ابزارهای دیگر برای کاربران را دارا می‌باشند. همچنین انبار داده‌ها ترکیب شده با فرایند پاک کردن داده‌ها ترکیب و یا تبدیل آنها به صورت بازه‌ای می‌باشد. با استفاده از انبار داده‌ها، کسب و کارها قادر به ارائه تحلیل‌های جامع، پیش‌بینی‌ها، و تحلیل‌های موقعیتی می‌باشند که در انتخاب بهترین استراتژی‌های تجاری به آنها کمک می‌کند. همچنین، انبار داده‌ها به صورت مداوم بروزرسانی می‌شوند و به صورت پیش‌بینی شده، به توسعه کسب و کار کمک می‌کنند. هر دو انبار داده و پایگاه داده، مجموعه‌ای از داده‌ها هستند. اما تفاوت اصلی بین این دو این است که پایگاه داده برای ذخیره و به روزرسانی داده‌ها به کار می‌رود، در حالی که انبار داده برای ذخیره و استخراج داده‌ها به کار می‌رود. یک پایگاه داده معمولاً برای ذخیره و مدیریت داده‌های یک سیستم کاربردی مورد استفاده قرار می‌گیرد، مانند یک سیستم مدیریت محتوا (CMS)، سیستم مدیریت انبارها (WMS)، یا سیستم مدیریت مشتریان (CRM). داده‌های ذخیره شده در پایگاه داده، معمولاً در قالب جداول با رابطه‌های بین آنها قرار می‌گیرند. از طرفی، انبار داده برای جمع‌آوری و تجمیع داده‌ها از منابع مختلف استفاده می‌شود و برای پشتیبانی از تصمیم‌گیری‌های کسب و کار و تحلیل داده‌ها طراحی شده است. انبار داده معمولاً داده‌ها را به صورت غیر قابل تغییر و در قالب داده‌های مسطح (flat) ذخیره می‌کند. این داده‌ها اغلب در قالب جداول بزرگ‌تر با فضای بیشتری برای مقایسه و تحلیل در اختیار کاربران قرار می‌گیرند.

**به طور خلاصه، پایگاه داده برای ذخیره و به روزرسانی داده‌ها استفاده می‌شود، در حالی که انبار داده برای استخراج و تحلیل داده‌ها طراحی شده است.**

## سوال سوم :

یکی از روش‌های یافتن داده‌های پرت استفاده از توزیع نرمال<sup>۵</sup> و percentile ها است. در مورد این روش تحقیق کرده و آن را توضیح دهید.

### پاسخ:

توزیع نرمال یا توزیع گاوسی یکی از مهم‌ترین توزیع‌های احتمالاتی است که در بسیاری از رشته‌های علوم، فناوری، مهندسی و اقتصاد مورد استفاده قرار می‌گیرد. توزیع نرمال، به صورت یک منحنی گوسی به نمایش درمی‌آید که شکل آن به صورت یک پیک نواری می‌باشد که در وسط دارای میانگین و در اطراف آن، به تدریج کاهش می‌یابد. اگر داده‌های ما از یک توزیع نرمال پیروی کنند، می‌توانیم از این توزیع برای تشخیص داده‌های پرت استفاده کنیم. یعنی اگر داده‌ای که در دست داریم، بسیار دور از میانگین (یا بسیار نزدیک به نقاط کمیته) باشد، به عنوان یک داده پرت شناخته می‌شود. در مورد percentile ها، می‌توان گفت که این مفهوم به معنای محل قرار گرفتن یک داده در میان مجموعه داده‌ها است. به عبارت دیگر، اگر داده مورد نظر ما در میان ۹۰ درصد بزرگترین داده‌ها باشد، به عنوان یک داده عادی و نه پرت شناخته می‌شود. اما اگر داده‌ای در میان ۱۰ درصد کوچکترین داده‌ها باشد، به عنوان یک داده پرت شناخته می‌شود. بنابراین، با استفاده از percentile ها و توزیع نرمال، می‌توانیم داده‌های پرت را شناسایی کرده و از آن‌ها صرف نظر کنیم و یا برای پردازش داده‌های ما، به طور صحیح به آن‌ها رسیدگی کنیم.

## سوال چهارم :

فرایند پاکسازی داده‌ها<sup>۶</sup> و نمایش داده‌ها<sup>۷</sup> را در نظر بگیرید:

(الف) فرایند پاکسازی داده‌ها را تعریف کنید.

(ب) اهمیت نمایش داده‌ها را بیان کنید و به یک مورد از چالش‌های آن اشاره کنید.

(ج) چرا پاکسازی داده‌ها یک فرایند مهم و پیشنیاز برای نمایش داده‌ها می‌باشد؟

### پاسخ :

#### الف :

فرایند پاکسازی داده‌ها شامل مجموعه‌ای از فعالیت‌هایی است که برای بررسی، تمیزکاری، و حذف داده‌های نامعتبر، ناهمخوانی، و پرت از داده‌های مورد نظر صورت می‌گیرد. برای پاکسازی داده‌ها، ابتدا باید داده‌های در دسترس را بررسی و ارزیابی کرده و سپس به شیوه‌ای که با توجه به نیاز مورد استفاده قرار می‌گیرد، آن‌ها را تمیز کرد. این فرایند شامل اصلاح خطاهای نگارشی و مفهومی، برطرف کردن مقادیر گم‌شده، تعویض و تبدیل فرمت، رفع تداخل‌های داده‌ها، و حذف داده‌های تکراری و پرت است. با پاکسازی داده‌ها، دقت و قابلیت اطمینان تحلیل داده‌ها بهبود می‌یابد.

#### ب :

نمایش داده‌ها یکی از مهم‌ترین مراحل در تحلیل داده‌هاست زیرا نحوه نمایش داده‌ها می‌تواند تأثیر قابل توجهی بر تفسیر و درک داده‌ها و در نتیجه تصمیم‌گیری‌های انجام شده داشته باشد. به عنوان مثال، اگر داده‌ها به شکل نامرتب و ناشیانه نمایش داده شوند، می‌تواند باعث ایجاد گمراهی و تداخل در تحلیل داده‌ها شود. از طرف دیگر، نمایش مناسب داده‌ها به

محققان کمک می‌کند تا بتوانند به راحتی پدیده‌های مهم در داده‌ها را شناسایی کرده و دقیق‌تر تحلیل کنند. یکی از چالش‌های نمایش داده‌ها، مسئله انتخاب روش نمایش مناسب برای داده‌هایی با حجم بالا است. انتخاب روش نمایش مناسب برای داده‌های حجیم به دلیل پیچیدگی و نیاز به بررسی مجموعه‌ای از متغیرها می‌تواند چالش برانگیز باشد. همچنین، در صورتی که نمایش داده‌ها نامناسب باشد، ممکن است از داده‌های مفیدی برای تحلیل چشم‌پوشی شود و در نتیجه تصمیم‌گیری‌های نادرستی اتخاذ شود.

## ج :

پاکسازی داده‌ها به عنوان یکی از مراحل اصلی پردازش داده‌ها، بسیار مهم و پیش نیاز برای نمایش داده‌ها است. در صورتی که داده‌ها در مرحله پاکسازی صحیح و کاملی را پیدا نکنند، این داده‌ها نمی‌توانند به درستی تحلیل و درک شوند. علاوه بر این، داده‌های پاک‌شده می‌توانند به عنوان ورودی مناسب برای مدل‌های یادگیری ماشین و الگوریتم‌های مختلف به کار رود. در مرحله نمایش داده‌ها، داده‌های پاکسازی شده باید به گونه‌ای نمایش داده شوند که بتوانند به راحتی تحلیل و درک شوند. این نمایش می‌تواند شامل چندین متغیر و ویژگی باشد که برای توضیحات بهتر و یکپارچه‌تر داده، به یکدیگر وصل شده‌اند. این نمایش باید برای کاربران قابل فهم و قابل استفاده باشد تا بتوانند به راحتی داده‌های مورد نیاز خود را پیدا کرده و از آنها استفاده کنند. یکی از چالش‌های نمایش داده‌ها، مدیریت حجم بزرگ داده‌ها است. در صورتی که داده‌های بزرگی باید نمایش داده شوند، باید از روش‌هایی مانند فشرده‌سازی و کاهش ابعاد استفاده کرد تا بتوان حجم داده‌ها را کاهش داد و از کارایی مناسبی برخوردار بود.

## سوال پنجم :

در یک آزمایشگاه ژنتیک مقدار فعالیت دو ژنوم مختلف مورد بررسی قرار گرفته و در ۱۰ بازه زمانی مختلف در به صورت زیر ثبت شده است:

Gen\time	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
G1	-3	5	8	-2	1	2	3	-5	10	-1
G2	9	20	16	8	2	10	-6	-15	25	-2

الف) با استفاده از معیار شباهت Cosine Similarity, Correlation, Mutual Information شباهت این دو ژن را مقایسه کنید.

ب) طبق نتایج هر معیار مشخص کنید آیا دو ژنوم از یکدیگر مستقل هستند یا خیر.

ج) آیا نتایج به دست آمده متفاوت است؟ اگر پاسخ مثبت است علت آن را توضیح دهید.

## پاسخ :

### الف :

برای شباهت کسینوسی :

$$G1.G2 = -27 + 100 + 128 - 16 + 2 + 20 - 18 + 75 + 250 + 2 = 516$$

$$||G1|| = (9 + 25 + 64 + 4 + 1 + 4 + 9 + 25 + 100 + 1)^{0.5} = 15.55$$

$$||G2|| = (81 + 400 + 256 + 64 + 4 + 100 + 36 + 225 + 625 + 4)^{0.5} = 42.36$$

$$\cos(G1, G2) = \frac{G1.G2}{||G1||*||G2||} = \frac{516}{15.55*42.36} = 0.78322$$

برای Correlation :

$$\begin{aligned} Corr(x, y) &= \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} \\ &= \frac{\langle x - \bar{x}, y - \bar{y} \rangle}{||x - \bar{x}|| ||y - \bar{y}||} \\ &= CosSim(x - \bar{x}, y - \bar{y}) \end{aligned}$$

میدانیم که پس ابتدا میانگین داده ها را محاسبه کرده و سپس داده ها را از آن ها کم میکنیم و با جداول جدید شباهت کسینوسی حساب میکنیم .

$$\overline{G1} = (-3 + 5 + 8 - 2 + 1 + 2 + 3 - 5 + 10 - 1)/10 = 1.8$$

$$\overline{G2} = (9 + 20 + 16 + 8 + 2 + 10 - 6 - 15 + 25 - 2)/10 = 6.7$$

$$G1' = (-4.8, 3.2, 6.2, -3.8, -0.8, 0.2, 1.2, -6.8, 8.2, -2.8)$$

$$G2' = (2.3, 13.3, 9.3, 1.3, -4.7, 3.3, -12.7, -21.7, 18.3, -8.7)$$

$$G1'.G2' = 395.4$$

$$||G1'|| = (209.6)^{0.5} = 14.47$$

$$||G2'|| = (1346.1)^{0.5} = 36.69$$

$$\cos(G1', G2') = \frac{G1'.G2'}{||G1'||*||G2'||} = \frac{395.4}{14.47*36.69} = 0.74476$$

برای Mutual information :

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} P(x, y) \log \frac{P(x, y)}{P(x)P(y)},$$

میدانیم فرمول به صورت زیر است پس برای هر خانه این را حساب کرده و جمع میکنیم. فرض میگیریم همه اعداد مثبت است زیرا لگاریتم منفی نداریم.

برای هر ۲۰ خانه باید مقادیر را حساب کرده و جمع کنیم .

جمع کل خانه های ردیف اول برابر است با : ۴۰

جمع کل خانه های ردیف دوم برابر است با : ۱۱۳

جمع کل خانه ها برابر است با : ۱۵۳

جمع ستون ها به ترتیب از اول تا اخر برابر است با :

3	35	20	9	12	3	10	24	25	12	sum
---	----	----	---	----	---	----	----	----	----	-----

حال با توجه به فرمول مقادیر برای هر خانه را حساب کرده و جمع میکنیم .

G1	-0.00038	-0.00038	0.00551	-0.00152	0.00068	-0.00255	0.00206	-0.00063	0.00252	0.00068
G2	0.00039	0.00453	-0.00465	0.00181	-0.00058	0.05428	-0.00174	0.00065	-0.00237	-0.00058

جمع خطا شباهت برابر است با : ۰/۰۵۷۷۳

ب :

برای دو معیار اول نشان میدهد دو ژنوم از یکدیگر مستقل نیستند. اما در روش سوم کاملاً مستقل اند

ج :

مورد سوم با دوتای اول فرق دارد دلیل آن ها یک تبدیل اعداد به کاملاً + است و دیگری اینکه تمامی موارد رو نسبت به هم مقایسه میکند و نقطه به نقطه نیست.

## سوال ششم :

دو مورد از روش های data preprocessing روش های aggregation و sampling هستند. این دو روش را توضیح داده و مزایا و معایب هر یک را بنویسید.

پاسخ :

Aggregation یا تجمیع، به عملیاتی گفته می شود که در آن داده ها بر اساس یک یا چند ویژگی (فیلد) مشترک گروه بندی شده و سپس برای هر گروه، یک مقدار خلاصه یا آماری محاسبه می شود. به عنوان مثال، میانگین، میانه، حداکثر، حداقل و تعداد داده ها برای هر گروه قابل محاسبه است. این روش برای خلاصه سازی داده های بزرگ و پیچیده و استفاده در تحلیل های آماری و تصمیم گیری بسیار کاربردی است. مزایای این روش شامل حفظ کیفیت داده ها، خلاصه سازی داده ها و محاسبه آمار دقیق برای تصمیم گیری های بعدی می باشد. از معایب آن می توان به ازدحام داده های پراکنده و از دست دادن جزئیات داده ها اشاره کرد.

Sampling یا نمونه گیری، به عملیاتی گفته می شود که در آن برای کاهش حجم داده، یک سری داده تصادفی از مجموعه داده های اولیه انتخاب می شوند و بر روی آن ها تحلیل های آماری اعمال می شوند. این روش به دلیل کاهش حجم داده ها و هزینه کاهش محاسبات بسیار کاربردی است. مزایای این روش شامل کاهش حجم داده ها، کاهش هزینه محاسباتی و تقلیل پیچیدگی داده ها می باشد. از معایب آن می توان به از دست رفتن برخی از داده های مهم و احتمال ناصحیح بودن نتایج اشاره کرد.

## سوال هفتم :

در رابطه با کاهش بعد تحقیق کرده و به سوالات زیر پاسخ بدهید.

الف) مفاهیم انتخاب ویژگی<sup>۸</sup>، استخراج ویژگی<sup>۹</sup> و مهندسی ویژگی<sup>۱۰</sup> را توضیح و تفاوت‌های بین آن‌ها را بیان کنید.

ب) الگوریتم‌های کاهش بعد به دو دسته خطی و غیرخطی تقسیم می‌شوند. تفاوت این دو دسته را توضیح داده و روش کار الگوریتم PCA از دسته خطی و الگوریتم t-sne از دسته غیرخطی را توضیح دهید.

## پاسخ :

می‌کنند. به طور کلی، می‌توانیم به سه نوع ویژگی در داده‌ها اشاره کنیم:

### انتخاب ویژگی (Feature Selection) :

این فرآیند به انتخاب یا حذف ویژگی‌های موجود در داده‌ها می‌پردازد. هدف اصلی این فرآیند، کاهش تعداد ویژگی‌های استفاده شده در مدل‌های یادگیری است. این کاهش می‌تواند بهبود کارایی مدل‌های یادگیری و کاهش پیچیدگی محاسباتی آن‌ها را داشته باشد.

### استخراج ویژگی (Feature Extraction) :

در این روش، با استفاده از روش‌های مختلف، ویژگی‌های جدیدی از داده‌ها استخراج می‌شوند که بیانگر خصوصیات مهم آن‌ها هستند. برخلاف انتخاب ویژگی، این روش به جای حذف ویژگی‌ها، ویژگی‌های جدیدی ایجاد می‌کند.

### مهندسی ویژگی (Feature Engineering) :

در این روش، ویژگی‌های موجود در داده‌ها تغییر یا بهبود می‌یابند تا بتوانند بهبود کارایی مدل‌های یادگیری را ایجاد کنند. به عبارت دیگر، در این روش، ویژگی‌های موجود به گونه‌ای تغییر داده می‌شوند که کارایی مدل‌های یادگیری بهبود یابد.

مزایا و معایب روش‌های انتخاب ویژگی، استخراج ویژگی و مهندسی ویژگی به شرح زیر است:

### انتخاب ویژگی:

#### مزایا:

کاهش پیچیدگی مدل: با حذف ویژگی‌های غیرضروری و کم اهمیت، می‌توان پیچیدگی مدل را کاهش داد و به دقت بیشتری دست یافت.

بهبود عملکرد مدل: با حذف ویژگی‌های نامربوط و بدون ارتباط با هدف، می‌توان بهبود عملکرد مدل را به دست آورد.

#### معایب:

اطلاعات از دست می‌رود: با حذف ویژگی‌هایی که ممکن است در برابر هدف مدل مهم باشند، اطلاعات مهمی از دست می‌رود. نیاز به دانش خاص: برای انتخاب ویژگی‌های مناسب، نیاز به دانش خاص در زمینه داده‌ها و مدل‌سازی دارید.

### استخراج ویژگی:

#### مزایا:

افزایش دقت: با استفاده از ویژگی‌های مناسب، می‌توان دقت مدل را افزایش داد.



افزایش سرعت: با کاهش تعداد ویژگی‌ها، می‌توان سرعت محاسبات را افزایش داد.

#### معایب:

پیچیدگی بیشتر: استخراج ویژگی‌های پیچیده، می‌تواند باعث افزایش پیچیدگی مدل شود. پردازش داده بیشتر: استخراج ویژگی‌های پیچیده، ممکن است به پردازش داده بیشتر نیاز داشته باشد.

#### مهندسی ویژگی :

#### مزایا:

افزایش دقت مدل: با ایجاد ویژگی‌های بهینه و جدید، دقت مدل‌های یادگیری بالا می‌رود و احتمال بدست آوردن پاسخ درست بیشتر می‌شود.

کاهش پیچیدگی مدل: با استفاده از ویژگی‌های بهتر و جدید، می‌توان پیچیدگی مدل‌های یادگیری را کاهش داد و باعث افزایش سرعت یادگیری می‌شود.

جلوگیری از بیش‌برازش: با استفاده از مهندسی ویژگی، می‌توان جلوگیری از بیش‌برازش کرد که در صورت وجود ویژگی‌های غیرمناسب ممکن است رخ دهد.

#### معایب:

نیاز به دانش کارشناسی: برای ایجاد ویژگی‌های بهتر، نیاز است که کارشناسان با دانش کافی در زمینه داده‌ها و مدل‌های یادگیری دارای تخصص و تجربه باشند.

زمان‌بر بودن: ایجاد ویژگی‌های جدید و بهتر زمان‌بر و هزینه‌بر است و نیاز به تحلیل دقیق و گاهی آزمایش تعداد زیادی از ویژگی‌ها دارد.

احتمال افزایش ابعاد داده: با افزودن ویژگی‌های جدید و بهینه، احتمال افزایش ابعاد داده و در نتیجه پیچیدگی محاسباتی افزایش می‌یابد.

#### ب:

الگوریتم‌های کاهش بعد خطی و غیر خطی با توجه به روش اعمال تغییر بر روی داده‌ها تفاوت دارند. در الگوریتم‌های کاهش بعد خطی، تغییرات بر روی داده‌ها به صورت خطی صورت می‌گیرد، به عبارت دیگر، خروجی به صورت ترکیب خطی از ورودی هاست. این الگوریتم‌ها مانند PCA، LDA و NMF از این دسته هستند. از طرف دیگر، الگوریتم‌های کاهش بعد غیر خطی، تغییرات بر روی داده‌ها به صورت غیرخطی اعمال می‌شود و خروجی به صورت ترکیب غیرخطی از ورودی هاست. این الگوریتم‌ها مانند t-SNE و Kernel PCA از این دسته هستند. با توجه به این تفاوت در روش اعمال تغییرات، الگوریتم‌های کاهش بعد خطی برای داده‌هایی که قابلیت توصیف خطی دارند مناسب هستند، بنابراین از این الگوریتم‌ها برای داده‌های توصیف شده توسط ویژگی‌های عددی استفاده می‌شود. از طرف دیگر، الگوریتم‌های کاهش بعد غیرخطی برای داده‌هایی که قابلیت توصیف خطی ندارند، مناسب هستند، بنابراین از این الگوریتم‌ها برای داده‌هایی که توسط ویژگی‌های غیر عددی توصیف می‌شوند استفاده می‌شود.

الگوریتم PCE (Principal Curve Estimation) یک الگوریتم کاهش بعد غیر خطی است که به عنوان یک روش مهم در آنالیز داده‌های تصویری و شناسایی الگو به کار می‌رود. این الگوریتم قادر است برای داده‌هایی که به شکل خم‌دار و معکوس S شکل هستند، یک منحنی برتر (principal curve) را بیابد. الگوریتم PCE با ایجاد یک منحنی برتر بین داده‌ها، با توجه به نزدیکی آن‌ها به منحنی، می‌تواند اطلاعات بیشتری از داده‌ها استخراج کند. در این الگوریتم، منحنی برتر به صورتی تعریف می‌شود که بیشترین توضیح‌دهی را برای داده‌های مشاهده شده داشته باشد. برای یافتن منحنی برتر، الگوریتم

PCE از روش اختلاف مربعات نقطه‌ای استفاده می‌کند و با استفاده از الگوریتم بهینه‌سازی نقطه مرکزی، منحنی برتر را به دست می‌آورد. به عنوان یک الگوریتم کاهش بعد غیر خطی، الگوریتم PCE می‌تواند مزایایی مانند استخراج ویژگی‌های پیچیده از داده‌ها و ارائه نمایش‌های سطح بالا از داده‌ها را داشته باشد. اما این الگوریتم دارای معایبی نیز می‌باشد که مهمترین آنها شامل پیچیدگی محاسباتی بالا و وابستگی به انتخاب اولیه منحنی برتر و پارامترهای مشخصه الگوریتم است.

الگوریتم t-SNE یا t-Distributed Stochastic Neighbor Embedding یک الگوریتم غیر خطی برای کاهش بعد داده‌های پیچیده است که در کاهش بعد داده‌های بسیار پیچیده و چگال، مثل تصاویر، استفاده می‌شود. t-SNE در واقع تلاش می‌کند نقاط پراکنده در فضای بعد بالا را به صورت خوشه‌ای در فضای کم بعدی قرار دهد. با استفاده از t-SNE، نقاطی که در فضای بعد بالا به هم نزدیک بوده‌اند، در فضای کم بعدی نیز به هم نزدیک خواهند بود. همچنین، نقاطی که در فضای بعد بالا دور از هم بوده‌اند، در فضای کم بعدی نیز دور از هم خواهند بود. به این ترتیب، t-SNE برای نمایش دادن الگوهای پیچیده و تفاوت‌های ریز در داده‌ها بسیار کارآمد است. در روش کار t-SNE، ابتدا یک ماتریس شباهت برای داده‌ها تعریف می‌شود. سپس، با استفاده از الگوریتم گرادیان کاهش مرتبه دوم، ماتریس شباهت در فضای بعد کمتر قرار می‌گیرد. در نهایت، با استفاده از یک الگوریتم بهینه‌سازی، ماتریس شباهت در فضای کم بعدی به خوشه‌های مختلف تقسیم می‌شود. مزیت اصلی t-SNE این است که قادر به نمایش دادن الگوهای پیچیده و تفاوت‌های ریز در داده‌ها است. با این حال، معایبی نیز دارد، از جمله:

t-SNE یک الگوریتم محاسباتی سنگین است و زمان بر است، به خصوص برای داده‌های بزرگ.

t-SNE یک الگوریتم پارامتری است و نتایج آن ممکن است به شدت تحت تاثیر پارامترهایی باشد که برای آن تعیین می‌شوند.

## سوال هشتم :

برای داده‌های عددی زیر نمودار جعبه<sup>۱۱</sup> را رسم کنید.

۲۷, ۳, ۱, ۲۹, ۲۷, ۷۰, ۲۶, ۳۳, ۲۷, ۳۶, ۴۹, ۲۵, ۳۹, ۲۸, ۴۱

## پاسخ :

ابتدا اعداد را به صورت مرتب شده مینویسیم .

۱ ۳ ۲۵ ۲۶ ۲۷ ۲۷ ۲۸ ۲۹ ۳۳ ۳۶ ۳۹ ۴۱ ۴۹ ۷۰

با توجه به این کمترین مقدار برابر ۱

چارک اول برابر ۲۶

میانه برابر ۲۸

چارک سوم برابر ۳۹

و بیشینه برابر ۷۰ است



## سوال نهم :

همانطور که می‌دانید، یکی از روش‌های مقایسه دو توزیع آماری استفاده از روش  $q-q$  plot است.

الف) نحوه کار این روش را توضیح دهید.

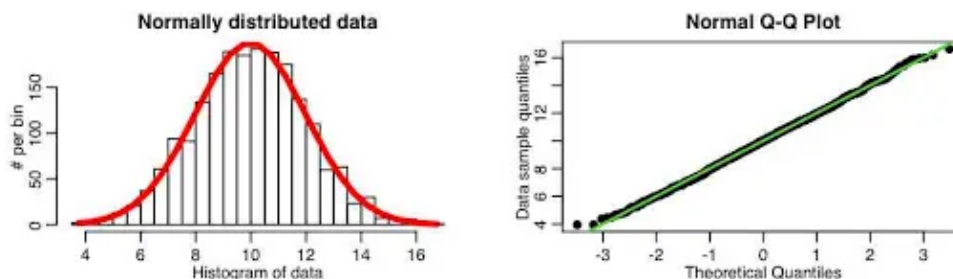
ب) نمودار  $q-q$  plot می‌تواند به شکل‌های متفاوتی نمایان شود: به طور مثال شبیه یک خط راست مورب. سه نوع از این شکل‌های متفاوت را بررسی کنید و تحلیل خود داده‌های توزیع‌های آماری ورودی به آن را بنویسید. به نظر شما از روی شکل  $q-q$  plot چه مواردی در مورد توزیع‌های آماری اولیه قابل استنتاج است؟

## پاسخ :

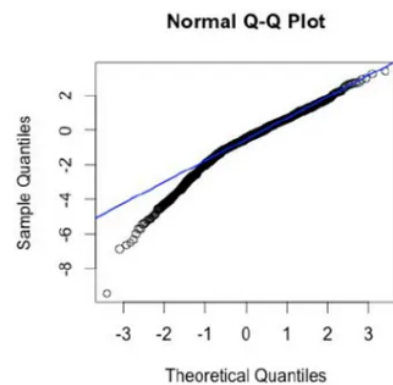
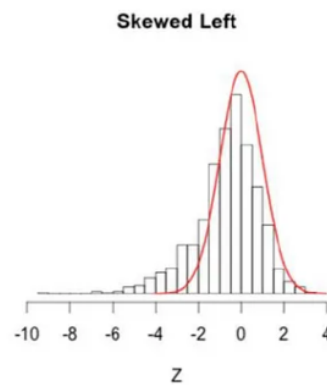
الف:

در این روش ما قسمت‌های مختلف دو دیتا ست را با هم مقایسه می‌کنیم به طوری که یک محور را برای یک دیتا ست و دیگری را برای دیتا ست دیگر در نظر می‌گیریم و حال داده‌های هر کدام را به صورت دوتایی  $(x, y)$  روی محور مختصات نشان می‌دهیم اگر دیتا ست‌ها برابر باشند روی خط  $x=y$  قرار می‌گیرد. با این روش می‌توان پیدا کرد که داده‌های ما از یک توزیع خاصی استفاده می‌کنند یا خیر و یا می‌توان داده‌های پرت را شناسایی کرد. و یا رابطه بین توزیع‌ها را که خطی اند و یا خیر.

ب :

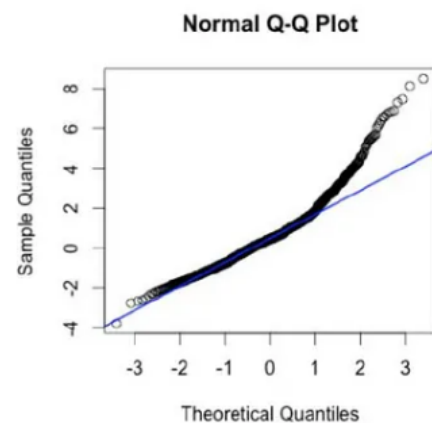
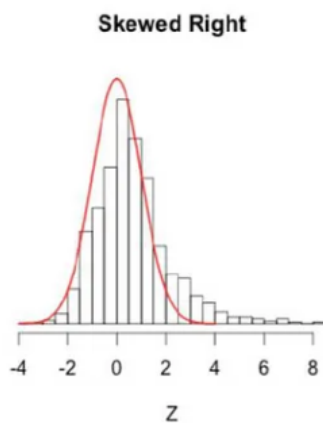


برای مثال در این تصویر می‌توان گفت که داده‌ها به صورت نرمال توزیع شده‌اند و مطابق قسمت الف روی خط  $x=y$  اند.



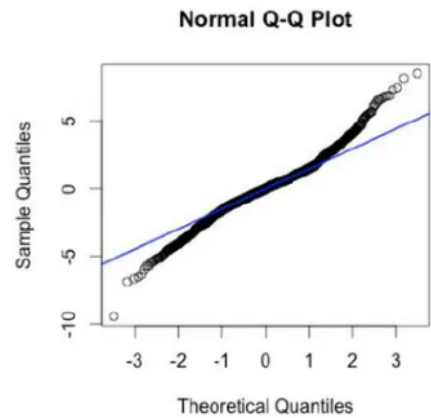
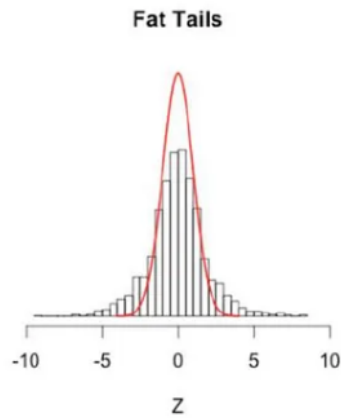
Left Skewed Q-Q plot for Normal Distribution

در این شکل نیز هم میتوان داده های پرت را تشخیص داده و هم اینکه مقایسه این دو توزیع که با توجه به q-q داده ها ابتدا روی توزیع مشخصی نیستند و سپس به سمت توزیع نرمال میروند.

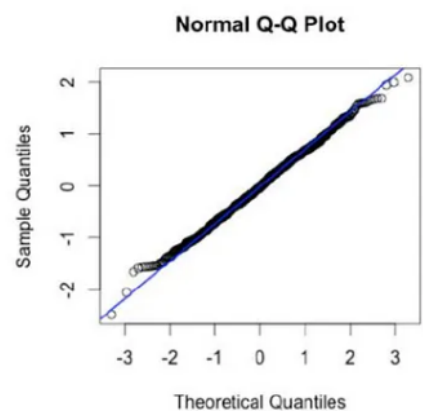
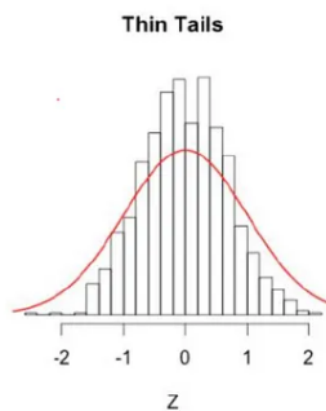


Right Skewed Q-Q plot for Normal Distribution

در این تصویر هم داده های انتها به صورت بالا هستند.



Fat-Tailed Q-Q plot for Normal Distribution



همچنین در این می توان مقایسه انواع توزیع ها با توزیع نرمال را دید و نحوه تشخیص داده هار روی  $q-q$  پس با توجه به نمودار میتوان پرت بودن داده و توزیع تقریبی و یا مورد انتظار خود را پیدا کرد.

## سوال دهم:

برای هر یک از روش های نرمال سازی زیر تحقیق کرده و بازه ی اعداد را مشخص کنید.

الف) نرمال سازی min-max

ب) نرمال سازی z-score

ج) نرمال سازی با مقیاس دهی<sup>۱۲</sup>

**پاسخ :**

**الف:**

با استفاده از فرمول روبرو تمامی اعداد را به بازه ۰ تا ۱ تبدیل میکند .

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

ب:

در کل برای درک بهتر اختلاف از میانگین است. داده ها را از میانگین  $\mu$  و انحراف معیار سیگما به مجموعه با میانگین ۰ و انحراف معیار ۱ تبدیل میکند . که فرمول آن مانند شکل زیر است. که رنج آن بین -۳ تا ۳ است.

$$Z = \frac{x - \mu}{\sigma}$$

ج :

بر اساس شیفیت دادن نقطه اعشار کار میکند . تمام اعداد را به بازه ۰ تا ۱ تبدیل میکند . و یا اگر منفی باشند -۱ تا ۱ تبدیل میکند. که ز توانی از ۱۰ است که بزرگترین داده را بین -۱ تا ۱ قرار می دهد.

$$U_i = \frac{V_i}{10^j}$$

**سوال یازدهم :**

با توجه به مقادیر ورودی  $X$  و مقادیر هدف  $Y$  می توان یک برازش خطی یا غیرخطی بر روی بسیاری از دادگان ها ایجاد کرد. با توجه به این مقادیر، به سوالات زیر پاسخ دهید.

$$X = [2, 4, 1, 3, 2, 6], \quad Y = [5, 6, 3, 6, 3, 10]$$

الف) روش محاسبه معادله نرمال<sup>۱۳</sup> را با استفاده از روش محاسبه مشتق جزئی باقی مانده<sup>۱۴</sup> کامل شرح دهید.

ب) یک برازش خطی ( $Y = \beta_1 X + \beta_0$ ) را برای این دادگان محاسبه کنید. (مقدار خطای برازش را نیز به دست آورید)

ج) یک برازش غیرخطی ( $Y = \beta_2 X^2 + \beta_1 X + \beta_0$ ) برای این دادگان محاسبه کنید. (مقدار خطای برازش را نیز به دست آورید)

پاسخ :

برای محاسبه معادله نرمال به صورت زیر عمل میکنیم :

ابتدا  $y$  را به صورت یک معادله فرض میکنیم :

$$y = \beta_0 + \beta_1 X$$

فرض کنیم باقی مانده ما به صورت زیر تعریف شود :

$$\epsilon = y - X\beta$$

حال تابع هزینه ما به صورت زیر تعریف می شود :

$$\begin{aligned} \epsilon^2 = (y - X\beta)^2 &\Rightarrow \|\epsilon^2\| = (y - X\beta)^T (y - X\beta) = (X\beta)^T X\beta - (X\beta)^T y - y^T X\beta + y^T y = \\ &(X\beta)^T X\beta - 2(X\beta)^T y + y^T y \end{aligned}$$

برای اینکه تابع هزینه کمترین مقدار باشد مشتق آن نسبت به ضریب بتا ما باید صفر شود تا مینیم در آن حوزه به دست آید.

$$\frac{\delta \|\epsilon^2\|}{\delta \beta} = 2X^T X\beta - 2X^T y = 0 \rightarrow \beta = (X^T X)^{-1} X^T y$$

ب :

$$X = \begin{pmatrix} 1 & 2 \\ 1 & 4 \\ 1 & 1 \\ 1 & 3 \\ 1 & 2 \\ 1 & 6 \end{pmatrix} \quad y = \begin{pmatrix} 5 \\ 6 \\ 3 \\ 6 \\ 3 \\ 10 \end{pmatrix} \quad X^T y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 4 & 1 & 3 & 2 & 6 \end{pmatrix} \begin{pmatrix} 5 \\ 6 \\ 3 \\ 6 \\ 3 \\ 10 \end{pmatrix} = \begin{pmatrix} 33 \\ 121 \end{pmatrix}$$

$$X^T X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 4 & 1 & 3 & 2 & 6 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 1 & 4 \\ 1 & 1 \\ 1 & 3 \\ 1 & 2 \\ 1 & 6 \end{pmatrix} = \begin{pmatrix} 6 & 18 \\ 18 & 70 \end{pmatrix}$$

$$B = (X^T X)^{-1} X^T y = \begin{pmatrix} 6 & 18 \\ 18 & 70 \end{pmatrix}^{-1} \begin{pmatrix} 33 \\ 121 \end{pmatrix} = \begin{pmatrix} \frac{11}{8} \\ \frac{11}{8} \end{pmatrix}$$

$$y = \frac{11}{8} + \frac{11}{8}x \quad \text{پس}$$

$$\begin{pmatrix} \frac{7}{8} \\ -\frac{7}{8} \\ \frac{1}{4} \\ \frac{1}{2} \\ \frac{1}{4} \\ \frac{3}{8} \end{pmatrix}$$

مقدار خطا نیز برابر است با که خطای متناظر هر نقطه است .

ج :

همانند قسمت قبل

$$X = \begin{pmatrix} 1 & 2 & 4 \\ 1 & 4 & 16 \\ 1 & 1 & 1 \\ 1 & 3 & 9 \\ 1 & 2 & 4 \\ 1 & 6 & 36 \end{pmatrix} \quad y = \begin{pmatrix} 5 \\ 6 \\ 3 \\ 6 \\ 3 \\ 10 \end{pmatrix} \quad X^T y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 4 & 1 & 3 & 2 & 6 \\ 4 & 16 & 1 & 9 & 4 & 36 \end{pmatrix} \begin{pmatrix} 5 \\ 6 \\ 3 \\ 6 \\ 3 \\ 10 \end{pmatrix} = \begin{pmatrix} 33 \\ 121 \\ 545 \end{pmatrix}$$

$$X^T X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 4 & 1 & 3 & 2 & 6 \\ 4 & 16 & 1 & 9 & 4 & 36 \end{pmatrix} \begin{pmatrix} 1 & 2 & 4 \\ 1 & 4 & 16 \\ 1 & 1 & 1 \\ 1 & 3 & 9 \\ 1 & 2 & 4 \\ 1 & 6 & 36 \end{pmatrix} = \begin{pmatrix} 6 & 18 & 70 \\ 18 & 70 & 324 \\ 70 & 324 & 1666 \end{pmatrix}$$

$$B = (X^T X)^{-1} X^T y = \begin{pmatrix} 6 & 18 & 70 \\ 18 & 70 & 324 \\ 70 & 324 & 1666 \end{pmatrix}^{-1} \begin{pmatrix} 33 \\ 121 \\ 545 \end{pmatrix} = \begin{pmatrix} \frac{1981}{890} \\ \frac{334}{445} \\ \frac{39}{445} \end{pmatrix}$$

پس  $y = \frac{1981}{890} + \frac{334}{445}x + \frac{39}{445}x^2$



$$= \begin{array}{r} 821 \\ 890 \\ - 561 \\ \hline 890 \\ - 57 \\ \hline 890 \\ 653 \\ 890 \\ - 959 \\ \hline 890 \\ 103 \\ 890 \end{array}$$

خطا نیز برابر است با      که خطای متناظر هر نقطه است .

## بخش عملی :

پاسخ سوالات عملی در فایل ژوپیتر است.

## سوال پنجم :

### پاسخ:

در یادگیری عمیق، Data Augmentation یکی از روش‌های مهم برای پیش‌پردازش داده‌ها است که به منظور افزایش حجم داده‌های آموزشی و افزایش دقت و عمومیت شبکه عصبی استفاده می‌شود. Data Augmentation به معنی تغییر شکل و محتوای تصاویر موجود در مجموعه داده است.

به طور کلی، Data Augmentation به صورت تصادفی روی تصاویر اعمال می‌شود تا تصاویر جدیدی ایجاد شوند. تبدیل‌هایی که در این روش استفاده می‌شود عبارتند از:

۱. Flip: تصاویر را به صورت افقی یا عمودی می‌چرخانند. برای مثال، تصویر ماشینی که در جهت چپ به راست قرار دارد، به جهت راست به چپ تغییر می‌کند.

۲. Rotation: تصویر به زوایای مختلفی چرخیده می‌شود.

۳. Zoom: تصویر بزرگ یا کوچکتر می‌شود.

۴. Translation: تصویر به بالا، پایین، چپ یا راست حرکت می‌کند.

۵. Brightness adjustment: شدت روشنایی تصویر افزایش یا کاهش می‌یابد.

۶. Color adjustment: رنگ‌های تصویر تغییر می‌کنند.

۷. Noise: نویز به تصاویر اضافه می‌شود.

این تبدیل‌ها می‌توانند به صورت ترکیبی هم برای تصاویر اعمال شوند. به عنوان مثال، تصویری می‌تواند به صورت تصادفی چرخیده، زوم شده و برای آن نویز اضافه شود. در عمل، روش data augmentation بیشتر برای داده‌های آموزش استفاده می‌شود تا به مدل کمک کند که با دیدن نمونه‌های جدید، بهتر و دقیق‌تر یاد بگیرد و بتواند با داده‌های جدید هم‌خوانی داشته باشد. در صورت استفاده از این روش برای داده‌های تست، داده‌های تست ما تغییر می‌کند و این می‌تواند به عملکرد مدل در پیش‌بینی‌های بعدی تاثیر بگذارد. بنابراین، برای داده‌های تست از داده‌های اصلی استفاده می‌شود تا نتایج دقیق‌تری دریافت شود.

## سوال ششم :

### پاسخ :

#### روش Upsampling :

در پردازش سیگنال و پردازش تصویر، یک روش افزایش ابعاد داده‌ها است. در این روش با افزایش تعداد نمونه‌های داده، از اطلاعات بیشتری استفاده می‌شود. برای افزایش تعداد نمونه‌ها در روش Upsampling، روش‌های مختلفی اعمال می‌شود که از جمله مهمترین آن‌ها عبارتند از:

۱. Linear Interpolation: در این روش، مقادیر بین دو نقطه موجود به صورت خطی تخمین زده می‌شوند. این روش بسیار سریع است و به راحتی قابل پیاده‌سازی است. با این حال، تنها برای داده‌های با توزیع خطی مناسب است و برای داده‌هایی که فاصله زیادی از یکدیگر دارند، مناسب نیست.

۲. Nearest Neighbor Interpolation: در این روش، برای هر نقطه‌ی جدید، نزدیک‌ترین نقطه‌ی موجود در داده‌ها به عنوان مقدار آن نقطه در نظر گرفته می‌شود. این روش در برخی موارد منجر به تصویری ناهموار می‌شود ولی سرعت بالایی دارد.

۳. Bilinear Interpolation: در این روش، برای تخمین مقدار نقاط جدید، از چهار نقطه اطراف آن نقطه استفاده می‌شود. مقدار نقطه‌ی جدید به صورت ترکیبی از این چهار نقطه حساب می‌شود.

این روش از جمله روش‌های معمول در کاهش نرخ خطا در مسائل تصویری مانند تشخیص چهره، تشخیص شیء و ... است.

#### روش DownSampling :

یک روش از داده برداری مصنوعی است که در آن تعداد نمونه‌های داده کاهش داده می‌شود. در این روش، برخی از نمونه‌ها با حذف تصادفی حذف می‌شوند تا تعداد کل نمونه‌ها در داده‌ها کاهش یابد.

این روش به طور خاص برای مواردی مفید است که تعداد نمونه‌های داده‌های مجموعه داده بسیار زیاد است و برای اجرای الگوریتم‌ها و مدل‌ها کندی ایجاد می‌شود. به عنوان مثال، اگر مجموعه داده شامل ۱۰۰,۰۰۰ نمونه باشد، می‌توان با استفاده از روش Downsampling، تعداد نمونه‌ها را به ۱۰,۰۰۰ یا حتی کمتر کاهش داد تا سرعت اجرای الگوریتم بالاتر رود. با این حال، یکی از مشکلات اصلی این روش، احتمال از دست دادن اطلاعات مهم است که ممکن است در نمونه‌های حذف شده وجود داشته باشد. بنابراین، انتخاب نمونه‌های حذف شده بسیار حساس است و باید با دقت انجام شود تا اطلاعات مهمی از دست نرود.

#### روش ترکیبی:

در واقع یک ترکیب از روش‌های upsampling و downsampling است. در این روش، ابتدا با استفاده از روش downsampling حجم داده‌های اصلی کاهش داده می‌شود و سپس با استفاده از روش upsampling حجم داده‌ها افزایش داده می‌شود. به این صورت که ابتدا داده‌های اصلی را با روش downsampling کاهش داده و سپس داده‌های کاهش یافته را با روش upsampling افزایش می‌دهیم تا حجم داده‌ها به حالت اولیه خود بازگردد.

استفاده از این روش به دلیل ترکیب دو روش upsampling و downsampling، می‌تواند باعث بهبود کیفیت داده‌ها و حفظ تعادل داده‌ها شود. اما در صورتی که از این روش برای کاهش حجم داده‌های بزرگ استفاده شود، ممکن است باعث افزایش زمان اجرا و نیاز به پردازش توزیع شده شود.

## سوال هفتم :

سوال هفتم - در مورد روش‌های smoteenn و smotetomek تحقیق کنید نحوه کار آن‌ها را توضیح دهید. وجه اشتراک این دو روش چیست؟

### پاسخ :

روش‌های SMOTEENN و SMOTETOMEK دو روش کاهش بیش‌برازش در داده‌های نامتوازن هستند که در ادامه توضیح داده خواهند شد:

SMOTETOMEK:

این روش از ترکیب دو روش over-sampling و under-sampling استفاده می‌کند.

در این روش، ابتدا با استفاده از روش SMOTE، داده‌های جدید برای کلاس‌های کمیاب ساخته می‌شود تا تعداد داده‌های هر کلاس به یک حد نزدیک شود. سپس، از روش Tomek Links استفاده می‌شود تا نقاطی که در فضای ویژگی‌ها به هم نزدیک هستند و در دو کلاس مختلف قرار دارند حذف شوند. در نهایت، باقیمانده داده‌های هر کلاس برای آموزش الگوریتم استفاده می‌شود.

SMOTEENN:

این روش نیز از ترکیب دو روش over-sampling و under-sampling استفاده می‌کند.

در این روش، ابتدا با استفاده از روش SMOTE، داده‌های جدید برای کلاس‌های کمیاب ساخته می‌شود تا تعداد داده‌های هر کلاس به یک حد نزدیک شود. سپس، نقاطی که بین دو کلاس مختلف در فضای ویژگی‌ها قرار دارند حذف می‌شوند. در نهایت، باقیمانده داده‌های هر کلاس برای آموزش الگوریتم استفاده می‌شود. به طور کلی، هدف این دو روش، کاهش بیش‌برازش در داده‌های نامتوازن و افزایش دقت الگوریتم‌های ماشینی است.

هر دو روش SMOTEENN و SMOTETomek برای مقابله با مشکل عدم تعادل کلاس در داده‌های دو کلاسه ایجاد شده‌اند. در هر دو روش ابتدا روش SMOTE برای افزایش تعداد نمونه‌های کلاس کمتر استفاده می‌شود، سپس روش Tomek link و ENN به ترتیب در روش SMOTETomek و SMOTEENN برای حذف نمونه‌های نامطلوب و افزایش تفکیک دهی بین کلاس‌ها مورد استفاده قرار می‌گیرند. در واقع هدف اصلی هر دو روش افزایش تعداد نمونه‌های کلاس کمتر و حذف نمونه‌های نامطلوب از داده‌هاست.